YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses.

Kim-Anh Lê Cao[1,2,#*], Florian Rohart[3,*], Leo McHugh[1#], Othmar Korn[3] and Christine A. Wells[3, 4$]

[1]Queensland Facility for Advanced Bioinformatics, The University of Queensland, St Lucia, 4072 Australia,
[2]Institute for Molecular Biology, The University of Queensland, St Lucia, 4072 Australia,
[3]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St Lucia, 4072 Australia
[4]The Institute for Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences Glasgow University, G12 8TA

[$]To whom correspondence should be addressed. Tel: +61733463853; Email c.wells@uq.edu.au

[#]Present Address:
[Leo McHugh], Immunexpress Inc., 1100 Dexter Avenue N, Suite 100, Seattle, WA 98109 USA
[Kim-Anh Lê Cao], The University of Queensland Diamantina Institute, Translational Research Institute, 37 Kent Street, Princess Alexandra Hospital, 4102 Australia

[*] These authors contributed equally to this work

ABSTRACT

Gene expression databases contain invaluable information about a range of cell states, but the question "Where is my gene of interest expressed?" remains one of the most difficult to systematically assess when relevant data is derived on different platforms. Barriers to integrating this data include disparities in data formats and scale, a lack of common identifiers, and the disproportionate contribution of platform to the 'batch effect'. There are few purpose-built cross-platform normalization strategies, and most of these fit data to an idealised data structure, which in turn may compromise gene expression comparisons between different platforms. YuGene addresses this gap by providing a simple transform that assigns a modified cumulative proportion value to each measurement, without losing essential underlying information on data distributions or experimental correlates. The Yugene transform is applied to individual samples and is suitable to apply to data with different distributions. Yugene is robust to combining datasets of different sizes, does not require global renormalization as new data is added, and does not require a common identifier. YuGene was benchmarked against commonly used normalization approaches, performing favourably in comparison to Quantile (RMA), z-score or rank methods. Implementation in the www.stemformatics.org resource provides users with expression queries across stem cell related datasets. Probe performance statistics including poorly performing (never expressed) probes, and examples of probes/genes expressed in a sample-restricted manner are provided. The YuGene software is implemented as an R package available from CRAN.

BACKGROUND

The pattern of expression across different cell or tissue types provides important information about the function or regulation of a gene. However experimental 'batch' from platform or laboratory sources remains a major barrier to systematically interrogating patterns across different datasets, a problem exemplified by the Microarray Quality Control Consortium (MAQC) studies (1,2). The variation imposed by different experimental sources frequently swamps genuine biological differences between samples, and has been addressed on a platform-by-platform basis using a number of approaches, which have been comprehensively reviewed elsewhere (3-5). To date there are few methods that address the integration of raw data from multiple technical sources, with a larger body of research looking at concordance or discordance of signatures from separate, platform-specific analyses (1,6,7).

The problem of data integration across multiple experiments has not been sufficiently addressed despite some obvious reasons for doing so. Increasing the number of samples improves statistical power for most analyses, and provides better insight into the heterogeneity of the biology being examined. Gene signatures descriptive of a developmental or disease process are more likely to be robust if these can be tested across many independently derived examples of the process under study. As well as increased reproducibility, the integration of data derived from different platforms may allow novel insight, from emergent patterns arising from the direct comparison of different experiments.

Most batch correction methodologies are designed to integrate data derived from a single platform, and address assumptions particular to that platform such as probe design or sequence composition. The most commonly used methods are summarized in Table 1. For example, Empirical Bayes methods can reduce technical batch effects on single platforms such as Affymetrix or Illumina microarrays. Popular implementations include COMBAT (5), and the quantile normalization strategies exemplified by RMA (8). These are applied across the entire group of samples being examined, which imposes difficulties when adding new samples to a series, or if the series is very large. The 'frozen' or fRMA method was developed to be less reliant on the group, but still uses information gathered from public data repositories to estimate ideal distributions for Affymetrix datasets (9,10). Despite these advances, these methods are not designed to integrate data generated on unrelated platforms.

While generally applied across different experimental series run on the same platform, COMBAT has been suggested as a means to integrate Affymetrix and Illumina microarray experiments (11), however there remains significant room for improvement, not least the problem that the group-normalized value is relative to the grouping of samples included, and is not sample-independent. These methods make assumptions about ideal distributions of the underlying data, and the extension of such approaches

across multiple expression platforms remains untested. Difficulties might be expected when comparing data that have fundamentally different distributions – as can arise when comparing different types of expression data derived from different platform technologies.

The problem of comparing different datasets has been addressed most successfully to date using rank-normalization approaches, for example by ranking Z-score values (12), as exemplified by the Tuberculosis database TBDB (13). Indeed, rank based approaches have been adopted generally for cross-platform analyses, but loss of relative expression can result in compressed 'fold changes', thus introducing potential biases in downstream comparative analyses (12). Once again, cross platform strategies most commonly compare post-analyzed data, rather than attempt to integrate primary information for direct comparisons. Nevertheless, systematic integration of primary data from multiple sources should be possible, despite the technical difficulties.

Proof-of-principle that robust tissue-specific and cell-type specific gene signatures can be derived across a range of public experimental datasets can be found in expression atlases such as ArrayExpress (14,15). Many of the most useful expression atlas approaches rely on data derived on a single technical platform, typically even avoiding direct comparison between different versions of the same platform. BioGPS (16) and the Gene Expression Commons (17) for example use data derived on the Affymetrix Genechip platform. Similarly, Pluritest (18) provides an example of a tool that benchmarks stem cell signatures generated on a specific version of the Illumina microarray platform. By relying heavily on a single platform, these large atlas projects face the very real danger of premature technical redundancy because of the rapid pace of technological change. While new atlas approaches are being scaled to interrogate genes across libraries of tissues or cells (19-21), it is arguably impossible to reproduce the variety of cell-states currently available in gene expression repositories such as ArrayExpress, or the Gene Expression Omnibus (GEO), particularly when one considers scarce patient-derived materials, or material from models that rely on knock-down or over-expression of target genes. This is exemplified by the sheer volume of available data: for example, there are now more than 4000 platforms listed in GEO for querying the human transcriptome, and these have generated in excess of 500,000 experimental samples (22). Therefore we need to continue to develop robust approaches for data integration, which can evaluate genuine biological patterns of gene expression across data derived from many different technical and experimental sources.

In the current study, we address two of the major limitations of current cross-platform normalization methods: the requirement for identical identifiers, which excludes or limits analysis from platforms with different probe-set identities; and the difficulties in merging datasets derived using different scales of measurements, such that datasets measuring different numbers of unique measurements, or generating different types of outputs cannot be easily compared. Our approach, 'YuGene', addresses the major

barriers identified above by accurately capturing relative intensity information using a modified cumulative proportion, which allows for a simple and robust method by which to directly compare different probes for a gene, and to compare samples within and between experiments. YuGene does not address all of the possible experimental confounders in a cross-platform meta-analysis, but it reduces the impact of technical batch on combined expression data. We demonstrate the utility of our approach to identify genes that are highly enriched in stem cell populations, using the www.stemformatics.org platform.

MATERIAL AND METHODS

1. The YuGene transform

YuGene uses the Cumulative Proportion transform. Let us denote by P_i the expression of the probesets on the chip, and by $P_{(i)}$ the expression of these same probes but in descending order, from highest to the lowest values (*i=1, …, p*).

$$Y_{(i)} = 1 - \frac{\sum_{j=1}^{i} P_{(j)}}{\sum_{j=1}^{p} P_{(j)}} = \frac{\sum_{j=1}^{p} P_{(j)} - \sum_{j=1}^{i} P_{(j)}}{S_p} = \frac{\sum_{j=i+1}^{p} P_{(j)}}{S_p} \text{ for } i = 1, \dots, p-1$$

$$Y_{(p)} = 0,$$

where $Y_{(i)}$ is the YuGene transformed value for probe(i), $P_{(i)}$ is the pre-processed raw value for probe(i)(see following section 2.a and S1), p is total number of probes on the array and $S_p = \sum_{j=1}^{p} P_{(j)}$. The output for each probe $Y_{(i)}$ is a value between zero (lowest expression) and close to one (highest expression) with intermediate values showing the proportion of the cumulative distribution captured up to and including the value for this probeset. Each sample is normalized without reference to other samples and values can be directly compared between samples without the need for renormalization. The YuGene algorithm is available as an R package on the Comprehensive R Archive Network (CRAN) (http://cran.r-project.org/web/packages/YuGene/index.html).

A 'tie' is considered when equivalent values occur in the raw data, for example $P_{(i)} = P_{(i+1)}$, and the same YuGene value is assigned to each member of the tie such that $Y_{(i)} = Y_{(i+1)} = 1 - \frac{\sum_{j=1}^{i} P_{(j)}}{\sum_{j=1}^{p} P_{(j)}}$.

2. Preprocessing steps

YuGene was benchmarked against raw data and other types of data transformation including quantile normalized (RMA), COMBAT, rank and z-score, detailed below.

a. *Normalization*. The preprocessing step involved a background correction and a log2 transformation of the raw values. Any 'raw' data values subsequently shown were log2 transformed and background corrected without further normalization. Pseudo code of the preprocessing steps is available in S1. Each dataset was preprocessed using an appropriate package from R/Bioconductor; Affymetrix arrays were handled with affy (23) (http://www.bioconductor.org/packages/release/bioc/html/affy.html) or oligo (24) (http://www.bioconductor.org/packages/release/bioc/html/oligo.html) and Illumina arrays were handled with lumi (http://www.bioconductor.org/packages/release/bioc/html/lumi.html) and the parameters

used are described in detail in S1. Quantile normalization was performed using the preprocessCore package (25). Where appropriate, COMBAT (26) was applied using the sva R package (http://www.bioconductor.org/packages/release/bioc/html/sva.html). Rank or z-score transformations were applied using the rank and scale functions (respectively) available in the R core base package (27). No subsequent analysis was run on individual Affymetrix probes, but all analyses were run at the summary probeset level. To avoid using platform-specific nomenclature, from this point expression values from an Affymetrix probeset are referred to simply as 'probe'

b. *Mapping*. Where a common identifier was required to compare the behavior of different normalization methods, the Ensembl gene identifier was used. When several probes from the same platform mapped to the same Ensembl gene ID, then the probe with the highest average expression was arbitrarily chosen to represent the gene.

c. *Detection threshold*. Where stated, the detection threshold was applied at the probe level on each raw dataset independently. The threshold was defined so that i) 20% of the probes were removed and ii) more than 1/3 of the samples had these probes under the detection threshold.

3. Datasets used

Three different types of microarray data were used to benchmark YuGene.

a. Simulated data

A first benchmark was performed on simulated data. YuGene transformation was compared to RMA (quantile normalization). The simulated data took its distribution from an Affymetrix study. Several data sets X were generated with n = 100 samples and p = 12,822 genes. In each of these data matrices, we simulated two types of cells (biological effect) and five types of batch (technical effect). Each condition (cell-type or batch effect) was randomly assigned to a sample. The expression of 12K genes were simulated, amongst which 10% had a true biological effect, (cell type) and 10% had a study (batch) effect. A possible overlap between these differentially expressed (DE) genes due to either a true biological effect and a batch effect was allowed. The code used to generate this data, and the simulated dataset itself is provided in Supplemental file S2-1 (Zip).

b. MAQC microarray data

The MicroArray Quality Control MAQC project was used as the community standard for assessment of intra and inter-platform variability of microarrays (1) and includes a number of reference RNA samples processed within and between labs (as technical replicates), and on a range of microarray platforms. We focused on the Affymetrix (HG-U133 Plus 2 arrays, 54K probes) and Illumina platforms (Human-6 arrays, 47K probes) using two reference RNAs (denoted 'reference A' for 100% of Stratagene's Universal Human Reference RNA, 'B' for 100% of Ambion's Human Brain Reference RNA). 3-4 technical replicates for

each reference RNA was obtained from 4-6 test sites. The data were obtained using the R packages MAQCsubsetAFX and MAQCsubsetILM.

c. Combined Stem Cell Microarray data from the Stemformatics resource.

The list of publicly available data sets used in our analyses is available in www.stemformatics.org and are described in Table 2.

4. Statistical analyses

a. Fold-Change

The fold-change was calculated as a difference between the mean log2 expression of two conditions.

b. Differential expression analysis.

The differential expression (DE) analysis was performed using a linear mixed model (nlme R package (28) with Maximum Likelihood estimation) with cell type (biological condition) as a fixed effect, and technical (or batch) as a random effect. For the specific case of MAQC, we included site, reference and platform as fixed effects and the technical or biological replicates as random effects. The p-values were obtained from an ANalysis Of VAriance (ANOVA) and were adjusted for multiple testing correction (BH<0.01 with the False Discovery Rate procedure from Hochberg & Benjamini (29)). The data were scaled prior to the analysis to obtain comparable and unbiased variances of the random effect between each tested method.

c. Principal Component Analysis (PCA).

PCA was performed on the first 2 principal components. Each data set was centered on the probes or genes and the graphs were obtained using the R package mixOmics (http://perso.math.univ-toulouse.fr/mixomics/).

d. Correlation

The concordance between technical replicates normalized with different methods was assessed using the Pearson's correlation coefficient between each pair of technical replicates. Additional details are provided in S3.

5. Implementation in Stemformatics

The YuGene transform was applied to all gene expression microarray datasets housed in Stemformatics. Probe-level YuGene expression values are stored in an in-memory lookup table for fast query access. YuGene "compare all" graphs are composed for each ENSEMBL gene, where all measurements mapping to that gene are displayed in a rank order from highest to lowest YuGene value. Each column represents

one measure on one sample, and is interactive, to direct a user to the original data profile (Data-set centric views are shown with platform-specific normalization, not including YuGene transformation). Samples are colored in the YuGene graph on a grey scale where each color block represents 0.1 of a unit on the YuGene scale, and under the YuGene graph as a barcode colored in red-scale on dataset ID. This allows users to quickly assess for dataset or platform biases in the YuGene patterns of expression. YuGene graphs are returned on the GeneSearch page, or can be navigated to via the Geneview menus.

RESULTS

1. YuGene preserves major features of data distributions

YuGene is a scaling method, in contrast to quantile normalization methods that fit the data to an idealized distribution, or rank transformations that ignore the distribution of the underlying data altogether. A comparison of the density distributions of these common strategies (Figure 1) is provided for a typical Affymetrix microarray sample. While both YuGene and the rank transformation resulted in a more linear probit function shape than z-score or RMA normalized data, YuGene retained more information about the underlying distribution than the ranked approach. YuGene was tolerant of different types of distributions, scaling but not imposing an identical fit. This is further shown in the density plots in S2 for a comparison of several real-world microarray studies taken from 6 different platforms (described in Table 2).

Scaling expression data using a YuGene transformation did result in compression of normalized values at the tails of the range of measurements taken. This is illustrated in Figure 2, where a linear relationship was observed with high concordance between data transformed using YuGene or with alternate normalization strategies (Pearson correlations >0.95, Figure 2), but YuGene-transformed data appeared compressed at the high and low values relative to the other methods. Measurements taken at the extreme of signal intensity frequently sit outside the linear range of the detection method used, and can be a major source of variation between manufacturers, platforms as well as laboratory sites. Reducing the impact of small changes at these extremes may therefore be an important consideration in reducing the false-positive call rate for comparisons of platform-dependent fluctuations. Importantly, the Fold-Change calculated on YuGene values were highly concordant with those calculated on Quantile normalized data, (see Figure 2(c-d)), and very little, if any compression of the fold-change values was observed.

We next assessed whether potentially inter-dependent measures could be robustly identified in YuGene transformed data (see S3). This would be expected to occur when multiple probes are provided for the same gene, or when two probes return an identical value, or if the expression of one probe is highly correlated with the expression of another. To do so, the probe-level correlations across technical

replicates measured at different MAQC sites were assessed. YuGene retained high correlations between technical replicates on both Affymetrix and Illumina microarray platforms (r>0.97 for Illumina and r>0.89 for Affymetrix platforms, both on detection-thresholded data), a trend confirmed for biological replicates from a number of additional datasets.

2. YuGene is tolerant of different sized datasets.

Allowing the inclusion of different identifiers will permit the inclusion of different sized datasets in a comparative series. For example, a standard Illumina microarray platform contains approximately 48,000 unique probes, whereas some Affymetrix platforms may measure 33,000 probes. If the number of measurements available for any single dataset impacts on the YuGene distribution then this may result in a different YuGene value being assigned to an equivalent measure. The resulting difference in YuGene values would reduce the reliability of any comparison made between those datasets.

To test this, a Wilcoxon test was used to test whether the distribution of YuGene values changed when a subset of probes was sampled from the same data set. This test was performed 200 times on random subsamplings, for different subset sizes. Figure 3 illustrates the severe impact that the number of probes has on the distribution of raw, quantile normalized and z-score transformed data, with a marked difference in Wilcoxon p-values observed even when 70% of the probes are retained. In contrast, Yugene transformed data was minimally impacted by changes to the number of probes sampled, indicating that the number of probes is not a major confounder when implementing YuGene. This property of YuGene is particularly useful for comparing platforms with large differences in the number of features.

3. YuGene reduces the 'batch' effect.

YuGene is applied to each individual sample, which means new data can be combined without the need for renormalization across the series. However most normalization procedures are applied across an experimental series in order to reduce technical (or batch) variation. We used Principal Component Analysis to provide a useful visualization for combined data treated with different normalization methods, in order to assess whether the proportion of variance in the dataset that could be attributable to technical or biological effects.

On the MAQC dataset (Figure 4) batch was a major driver of variation across all of the normalization methods, but only YuGene transformed data notably emphasized biological effect on the first component (47.4% of the total explained variance between brain vs universal reference RNA), and reduced the impact of laboratory site on the variance structure of the data (42.1%of the variance on the second component). Platform was the biggest driver of the first component for quantile normalized (50.1% of the total variance) or Z-score (48.7%) with separation between the two biological sources on the second

component (42% for quantile or Z-score). While the first two components in the COMBAT data explained a large fraction of the total variance (92.3%), it poorly resolved the biological source. Component 1 could be attributed to a 'site' batch effect that in combination with component 2 could partially explain the biological sources.

We further assessed the ability of YuGene to reduce the contribution of batch by combining several stem cell experiments, generated across multiple platforms in different laboratories, and sourced from the Stemformatics database (described in Table 2). Figure 5 (and S7) shows clear segregation between fibroblast and pluripotent stem cell types with all the normalization approaches, except for COMBAT, which required two PCA components to resolve the major biological clusters.

Thus YuGene was effective at reducing the impact of platform and laboratory source in multiple situations. In both scenarios – the highly technically controlled experiments (MAQC) and the merger of multiple public datasets - YuGene transformed data reduced experimental (batch) variability without loss of genuine biological variance.

4. YuGene transformed data is suitable for differential expression analysis

A linear mixed model was used to discern the accuracy of a differential expression analysis, using data simulated to have varying noise or batch effects. Cell type was treated as a fixed effect and batch as a random effect. Significance was assessed using a false discovery rate of 1%. As can be seen from Table S4, YuGene and Quantile normalized data returned similar percentages of Type I and Type II errors, with YuGene transformed data resulting in fewer Type I errors, but slightly higher false-negative rate than Quantile normalized data. Batch was easily dealt with in all cases, and unsurprisingly neither method worked well when the noise between technical replicates was modeled to be high.

Given the relative consistency of the two normalization methods on simulated data, concordance was assessed across the 18 MAQC samples, where batch included six laboratory sites and two microarray platforms (Affymetrix and Illumina). Differences were assessed between two biological sample types – A: the Universal human reference vs. B: human Brain. The model included site, reference and platform as fixed effects and the technical or biological replicates as random effects. Table 3 (and Figure S5) shows a high degree of overlap of differentially expressed genes was identified in YuGene or quantile normalized data. YuGene identified 6061 differentially expressed genes, of these 17% (1032 genes) could be attributed to a biological effect alone. Although more differentially expressed genes were predicted in the Quantile or Z-score normalized data, the technical effects accounted for a higher proportion.

Following the analysis on the MAQC technical replicates, the next step was to perform a differential analysis across several stem cell datasets described in Table 2. A linear mixed model was used, but the experimental source was treated as a random effect to allow us to quantify the amount of variance due to

the batch effect. The Venn Diagram in Figure 6 illustrates the overlap between the gene sets identified by each transformation, and the variance due to the batch effect was summarized in the boxplots. This was significantly less in the YuGene transformed data than in Quantile (one-sided t-test, p-value = 6.261e-8 and 4.544e-11 for (b) and (c)). See also S6 for more details.

In all three scenarios, YuGene transformed data reduced the contribution of technical variance, and returned fewer false positives than the other normalization strategies.

To further assess the influence of the YuGene compression of the high values, differential expression analyses were performed on the Guenther data set (30) (See Table 2) between the 17 hESC and 22 hiPSC cell lines for both YuGene and Quantile transformations. The large proportion of genes declared as differentially expressed by both transformations indicated that the YuGene compression of high values did not seem to affect the differential expression analysis results (see S7).

5. Application of YuGene to the Stemformatics database

The original motivation for development of the YuGene transform was the desire of Stemformatics.org users to query the behavior of an individual gene across multiple datasets. This frames the "where is my gene of interest expressed" question, one that is repeatedly requested by biologists wishing to ascertain whether patterns derived from their own data or observations can be recapitulated across the multiple unrelated datasets housed in Stemformatics (31).

Stemformatics currently houses fifty-four different data types, including twenty-eight different microarray chip-types that can be summarized into eight 'platforms' servicing mouse or human stem cell data (Table 4). Probes whose behavior is poor compared to the probeset to which they belong are flagged to allow users to remove these from downstream analyses. For example, Table 4 lists the % of probes on each platform that always sit 2 fold below the expression of other probes mapping to the same gene, and that are always expressed below the 3rd quartile of any probe. These might be considered to be poorly hybridizing probes, or probes directed to a transcript that is not commonly expressed and not seen in the samples that we have survey. Probes that do not vary in YuGene profile across the Stemformatics samples are excellent candidates for housekeeping controls in subsequent validation experiments. An example is given in Figure 7a, for the housekeeping gene ACTB. Probes that are positively skewed are expressed at a high level are considered to be commonly expressed across the majority of Stemformatics samples, likewise probes that are negatively skewed are considered to have a restricted pattern of expression, and may be cell-type specific. An example of the former is POU5F1 (Figure 7b), and the latter is illustrated by Figure 7c (DNMT3L).

DISCUSSION

The motivation for YuGene was to address a common request from Stemformatics.org users who wished to query “where is my gene of interest expressed”. Although this query may seem relatively trivial in the face of more sophisticated analyses of global gene expression, it is an important one that is surprisingly under-developed given the spectrum of gene expression databases that are in the public and private domain. The implementation of YuGene in Stemformatics provides a simple overview of every sample, which we further rank in decreasing order of YuGene value. This type of visualisation provides an intuitive way for stem cell researchers to assess groups of samples that may share common patterns of expression (for example, high levels of the pluripotency transcription factors POU5F1 in hESC and iPSC datasets but not differentiated cells shown in Figure 9), identify potential housekeeping genes which do not vary across the samples of interest, or find genes with a discrete expression pattern. By coloring the samples according to where they sit in the YuGene scale, small differences in rank in the tail ends of the YuGene distribution are minimized. By providing information back for all probes mapping to any Ensembl Gene, patterns that are dependent on just a single type of measurement are intuitive and obvious. An alternate approach would be to implement a visualization that highlights genes that show relative changes in expression from a common median (as given in the Gene Expression BarCode(32) project), but this makes assumptions about the comparative distributions of samples in Stemformatics that may not hold. As with any bioinformatics analysis, the YuGene transform does not remove the necessity to validation predictions made from cross-platform analysis. However by examining patterns that are reproducible across many datasets, the focus of this validation can be on high-confidence targets.

Advantages of the YuGene Transform

(i) No need for common probe identifiers. Most normalization strategies impose the requirement of common identifiers, which may necessitate the building of a minimal convergent dataset. Even when mapping probes from the most popular array platforms to a common annotation such as the same reference gene, the overlap between any two platforms was in the order of 60 - 70% (33,34). YuGene therefore preserves considerable information when combining multiple datasets. Retaining unique identifiers provides opportunities to systematically evaluate concordant and divergent data at a probe level within a platform, and between platforms. As can be seen in Table 4, each platform contains a small percentage of poorly performing probes, whose expression was flagged as consistently low and discordant with its associated probeset. The expression of most probes, regardless of platform, was normally distributed – consistent with the assumptions that a large proportion of genes are expressed in a large proportion of cell types/samples. Likewise, relatively few probes demonstrated highly skewed patterns of expression. Evaluation of probe performance across many samples, and concordance within an experimental series, provides an important tool for users to filter out probes that may be uninformative

or exaggerate platform differences. This is only possible if every measurement can be retained in the normalization process.

(ii) Stand-alone single-sample normalization. YuGene does not require renormalization every time a new sample is added to the experimental series, thus simplifying workflows and protecting the integrity of existing analyses, which may be otherwise affected by renormalization. This has been previously recognized as an important advantage for optimizing cross-experimental analyses (9,10).

(iii) Keep the relationships between probes. YuGene retains the correlation structure between probes, and between samples for a single dataset. While some downstream analyses may assume independence between the measurements taken for a given dataset (for example, gene set enrichment analyses (35)), other analysis approaches exploit this correlation structure (for example transcript splicing indices or eQTL analysis (36-38)). We therefore argue that this should be explicitly examined, and the probe-correlation structure removed as required. Further, YuGene does not force each sample into the same ideal distribution, but retains information about the underlying (raw) data structure. YuGene quickly and efficiently rescales data of varying library sizes or with large differences in probeset numbers, without compromising the relative distribution of the measurements. This observation could be particularly useful when comparing RNAseq libraries of varying depth or when assessing specialist libraries that target functional RNA subsets (miRNAs vs other noncoding RNA subsets for example).

(iv) Reduce technical variation, keep genuine biological differences. Differential expression analysis across platforms requires evaluation of possible technical confounders, regardless of the normalization method used. In our hands, using simulated data, or technically controlled MAQC data, or data generated from the combined Stemformatics.org website, the YuGene transform returned high quality differential expression predictions, and generally had reduced type I error compared to the other methods. By including technical sources of variation in our linear models, we were able to show a reduction in genes assigned as differentially expressed in the ‘noise’ class. We also showed using PCA that YuGene transformed data reduced the variance imposed by platform or batch, in contrast to the other transformations that we benchmarked against. We speculate that the compression of data at the tails of a YuGene scale may contribute to this biological stability, particularly if differences in the absolute linear range of measurements taken on different platforms were a major contributor to small variation in these tails.

(v) Permit different platforms to have different linear ranges. Each platform generates data with a range of values that is highly dependent on experimental parameters such as probe deposition, the flurophore or colormetric assay used, the scanner resolution, and site-site variation is scanner settings (39). A gene may be measured as highly expressed in two experiments but may be assigned two different values due

to technical differences in the platforms, making a direct comparison impossible to perform. Certainly all microarray hardware has an upper and lower linear range of reliable measurements, but although the field is familiar with thresholding data on detection limits, it is relatively uncommon to threshold data that may have exceeded the upper limits of the detection spectrum (40). By rescaling data to a common range, YuGene does compress the values that sit at the high or low extremes of a distribution. However, we have shown that this compression does not seem to affect the results of the subsequent statistical analysis. The compression of the expression values may in fact reduce false positives due to platform differences in scale, providing a more meaningful direct comparison of these extreme measurements.

(vi) Speed and ease of implementation. YuGene is computationally quick and easy to implement, requires implementation only once per sample, and scales across large series of experiments so is suitable for database implementation.

Caveats and cautions when using YuGene

YuGene rescales data, and this results in a compression of values that sit at the very extreme of a distribution. YuGene may be adversely impacted by the presence of a high number of undetected probes. We observed in some cases that the correlation between technical replicates was improved when undetected probes were removed prior to applying the YuGene transform. This was particularly apparent for Illumina microarray data (Figure S3), but provided no apparent benefit for others (Table S3).

The normal caveats regarding interpretation of downstream analyses across multiple independently generated samples must also apply to YuGene transformed data. No transformation can correct poor experimental design, inappropriate comparisons, or major differences in the handling of the biological processes being evaluated. Although YuGene was effective at reducing technical artifact, it was not able to completely remove platform-generated batch effects that may mask genuine biological differences, meaning that some differentially expressed calls might appear as a result of a batch effect. However, YuGene was able to remove the experimental bias previously reported in a meta-analysis of several different stem cell experiments (41), and demonstrates that robust data transformation prior to analysis (YuGene or Quantile in this instance) will reduce the influence of the technical sources of variation on downstream data analysis. YuGene was also effective at minimizing type I error in our simulated datasets, but did return higher type II errors as a result.

CONCLUSION

YuGene is a simple data transformation that is applied to individual samples, which allows for the serial inclusion of new samples in a database. YuGene is a variant of a cumulative proportion measure and enables the comparison of relative gene expression measured across multiple platforms and obtained

from multiple datasets. It provides the means to compare patterns of gene expression in samples that might share biological properties, but be measured on using different technologies. The transform performs favorably to alternate, commonly used normalization methods and has been implemented across 1928 samples from 9 platforms in the http://www.Stemformatics.org database.

SUPPLEMENTARY DATA

Supplementary Data includes expanded benchmarking of YuGene, code and data for simulated datasets, and are available online at Genomics.

ACKNOWLEDGEMENT

The authors wish to thank Rowland Mosbergen and Isaac Englart for implementation of YuGene in Stemformatics.

FUNDING

This research was supported by Australian Research Council project funding DP130100777 to CAW and KALC, and by the ARC Stem Cells Australia initiative to CAW. CAW is supported by a Queensland Government Smart Futures Fellowship. KALC is supported, in part, by the Wound Management Innovation CRC (established and supported under the Australian Government's Cooperative Research Centres Program).

Table 1: A brief comparison of commonly used normalization methods and their application to cross-platform data transformation.

Method	Unique ID	Single-array	Distribution assumption	Platform	R package
Quantile normalization (RMA)	✗	✗	No	Applicable to any data series	`affy`(23)
fRMA	x	✓	No	Implemented for Affymetrix, but can be more widely adaptable	`fRMA` (9,10)
Rank approach	✓	✓	No	Applicable to any data series	`r-base-core`(27)
Z-score	✓	✓	Gaussian	Applicable to any data series	`r-base-core`(27)
SCAN	✓	✓	Gaussian	Currently implemented for Affymetrix or Agilent microarray platforms, but can be more widely adaptable.	`SCAN.UPC` (7)
COMBAT	✗	✗	Gaussian	Applicable to any data series	`sva` (26)
YuGene	✓	✓	No	Applicable to any data series	`Yugene (current)`

Table 2. Description of the microarray data sets for the different platforms. Original accession numbers of each experimental series is provided, together with the Stemformatics identifier. When combined, the data set includes 12822 common genes (4879 after removing probes under the detection threshold). hESC is human Embryonic Stem Cell; iPSC is induced Pluripotent Stem Cell.

Database & Accession number (Stemformatics dataset ID)	Platform	Study (Author, reference)	Sample class		
			Fibroblast	hESC	iPSC
GEO: GSE20033 (5018)	Affymetrix HG-U133_Plus_2	Jia(42)	0	4	11
GEO: GSE23402 (5025)	Affymetrix HG-U133_Plus_2	Guenther(30)	3	17	22
GEO: GSE12390 (5027)	Affymetrix HG-U133_Plus_2	Maherali(43)	3	3	15
Array Express E-GEOD-14897 (6006)	Affymetrix HG-U133_Plus_2	Si-Tayeb(44)	3	0	3
GEO: GSE9709 (6161)	Affymetrix HG-U133_Plus_2	Masaki(45)	2	0	8

GEO: GSE9832 (6162)	Affymetrix HG-U133_Plus_2	Park(46)	6	1	8
GEO: GSE16093 (6165)	Affymetrix HG-U133_Plus_2	Kim(47)	0	1	3
GEO: GSE16654 (6166)	Affymetrix HG-U133_Plus_2	Chin(48)	1	1	4
GEO: GSE28970 (5002)	Affymetrix HT-HG-U133A	Bock(49)	6	20	12
GEO: GSE25673 (5012)	Affymetrix HuGene-1 0-ST v1	Brennand(50)	0	0	23
GEO: GSE36648 (6092)	Affymetrix HuGene-1 0-ST v1	Andrade(51)	3	6	15
GEO: GSE20532	Illumina HumanRef-8 V3	Zaheres(52)	0	4	6
GEO: GSE35347 (5036)	Illumina HumanHT-12 V4	Nayler(53)	6	3	12
ArrayExpress: E-MTAB-1040 (6072)	Illumina HumanHT-12 V4	Vitale(54)	8	3	18
Sample Total			**41**	**63**	**160**

Table 3. YuGene reduces the impact of site or platform on differential expression calls in MAQC data. The number of ENSEMBL genes called as significantly different between groups (adjusted p-value with BH< 0.01), and the impact of combining batch (site effect) or platform (Affymetrix and Illumina data) on the comparison of two different RNA samples (A universal reference RNA and B Brain RNA).

n = 18	Raw data (total probe number = 7389)	Quantile (total probe number = 7389)	Z-score (total probe number = 7389)	COMBAT (total probe number = 7389)	YuGene (total probe number = 7389)
Reference effect: A vs. B	5133	6333	6075	5013	6061
Site effect: 6 sites	7091	1718	2429	6980	1635
Platform effect: Affymetrix vs. Illumina	4852	6125	6081	1	5859

Table 4. Descriptive statistics of probe performance across Stemformatics datasets. The total number of probes includes only those probes that map to an ENSEMBL gene using the Stemformatics mapping pipeline. Discordant probes were defined where the individual probe returned a value at least 2X below the mean of all probes mapping to that gene, as well as 2X below the platform median. Probes were determined to be distributed without significant skew (Expected) if abs(Quartile measurement of skewness) < 0.2. Note that these numbers are entirely dependent on the number and type of samples evaluated, and are likely to change with the addition of more data.

Microarray Platform	Probe characteristics					# Samples evaluated
	Total Number Probes	discordant	Expected distribution	Extreme positive skewness > 0.5	Extreme negative skewness < -0.5	
Affymetrix HG-U133 Plus2	54,675	3,988 (7.3 %)	33,905 (62%)	1,211 (2.2%)	713 (1.3%)	565
Affymetrix Human Gene 1.0 ST V1	33,297	1,080 (3.2 %)	22,051 (66.2%)	384 (1.2%)	521 (1.6%)	234
Affymetrix Human Exon 1.0 ST V2	22,011	140 (0.6 %)	10,701 (48.6%)	1,195 (5.4%)	950 (4.3%)	36
Illumina Human HT-12 V4	47,323	2,608 (5.5 %)	29,472 (62.3%)	538 (1.1%)	895 (1.9%)	237
Illumina Human WG-6 V2	48,702	1,439 (3.0 %)	28,501 (58.5%)	868 (1.8%)	962 (2%)	90
Affymetrix Mouse 430 2.0	45,101	2,433 (5.4 %)	28,797 (63.9%)	1,547 (3.4%)	438 (1%)	306
Affymetrix Mouse Gene 1.0 ST V1	35,557	771 (2.2 %)	22,699 (63.8%)	601 (1.7%)	668 (1.9%)	156
Illumina Mouse Ref-8 V2	25,697	898 (3.5 %)	15,352 (59.7%)	767 (3%)	995 (3.9%)	149
Illumina Mouse WG-6 V2	45,281	3,014 (6.7 %)	29,503 (65.2%)	436 (1%)	812 (1.8%)	155

Figure 1. Evaluation of data transformation on data distributions: Density distributions are given for the same microarray sample (Fibroblast sample from Bock dataset (49)). (a) raw, (b) quantile normalized, (c) YuGene transformation, (d) rank of the data and (e) z-score transformation. The figures are obtained with a kernel density estimation. The values for two probes mapping to PAX6 (red) and CAT (blue) genes are indicated to provide relative scales and the median is indicated as a vertical line.

Figure 2. Correlation between replicates and fold change between biological groups. First row: Scatter plots of (a) raw vs Quantile normalized data and (b) raw vs YuGene transformed data, on one Fibroblast sample. Second row: Scatterplots comparing the Fold-change calculated between hiPSC and hESC (panel c) or hESC and fibroblasts (panel d) from Guenther dataset (30). Sample numbers provided in Table 2. X-axis is the Fold-Change calculated from Quantile normalised data, Y-axis is the Fold Change calculated from YuGene transformed data. The fold change is highly concordant between the two data transformations (Pearson correlations given for each comparison).

Figure 3. Assessment of dataset size on distribution of transformed data for: (a) raw, (b) quantile normalization, (c) YuGene transformation and (d) Z-score. Each series represents the proportion of probes randomly sampled prior to normalization (from 99%-1%). The y-axis provides the p-value of a Wilcoxon test, compared to the distribution of the full data set. The boxplots represent the summary of 200 random samplings from the data. Original data taken from Bock fibroblasts (n=6) sampled as part of a reference map of stem cells.

Figure 4. PCA of combined samples, showing the first two principal components for: (a) the raw MAQC data, (b) Quantile normalization, (c) YuGene transformation (d) Z-score and (e) ComBat batch correction. Six technical replicates from 6 different laboratories were assessed for Affymetrix HGU133plus2 microarray platform (open symbols), and 3 technical replicates from 3 different laboratories were assessed for Illumina REF6 human bead array (filled symbols). Sample A (blue) was universal human reference RNA and Sample B (green) was human brain RNA. The percentage of explained variance is provided for each component.

Figure 5. Clustering of samples combined from different stem cell data sets (data sets from Stemformatics described in Table 2). PCA plots for (a) raw data, (b) quantile normalization, (c) Yugene transformation and (d) Z-score are displayed for the first two principal components.

Figure 6. Comparison of DEG calls using different transformations, and assessment of the contribution of batch. Data sets from Stemformatics described in Table 2. A linear mixed model was performed on each transformation (raw, Quantile, YuGene) to identify DEG while considering the experiment/batch effect as a random effect. (a) is the Venn diagram showing the concordance of the DEG between all transformations (BH <0.01) and (b) the variance due to the batch effect on the common 3259 DEG and (c) specific DEG for each method (798, 45 and 39 genes for raw, Quantile and YuGene resp.).

Figure 7. Implementation in the Stemformatics.org resource: the x-axis of all panels represents the samples, ranked from highest to lowest YuGene value. The barcode under the x-axis color codes the datasets that each sample is derived from. Each sample is a narrow bar, where the height of the bar indicates its YuGene value (y-axis = YuGene values). The values are ordered from highest (most abundant) to lowest (least abundant) and colored according to a grey scale that groups them by tenth of 1. Panel A/ Human ACTB provides an example of a ubiquitously expressed gene, with YuGene value of >0.9 in most samples. Panel B/ Human POU5F1 provides an example of a gene with a normal distribution across all datasets in the Stemformatics expression database. Panel C/ Human DNMT3L provides an example of a highly restricted pattern of expression, with YuGene value>0.9 only found in very early embryos, and YuGene value>0.5 only found in pluripotent stem cells.

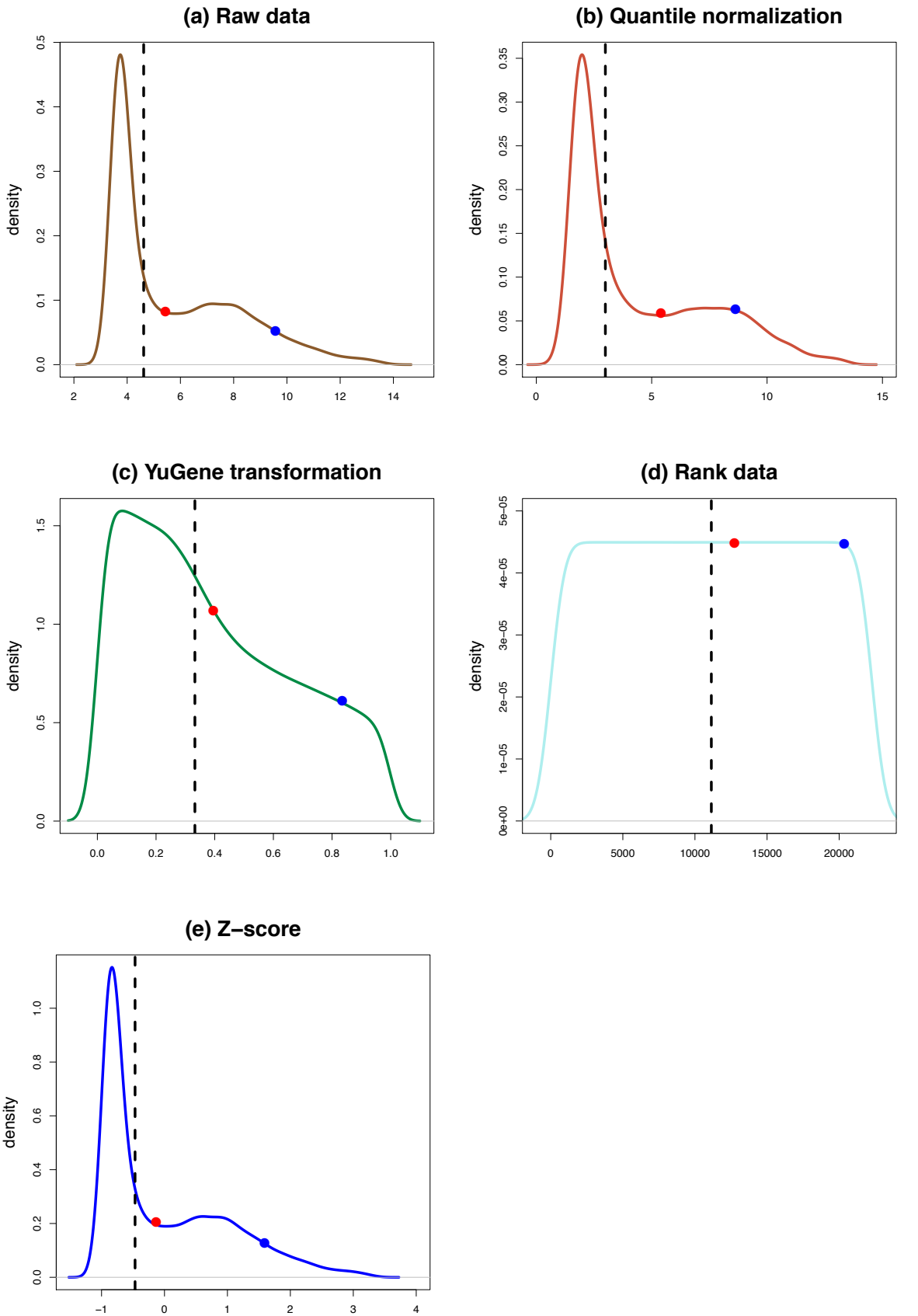
References

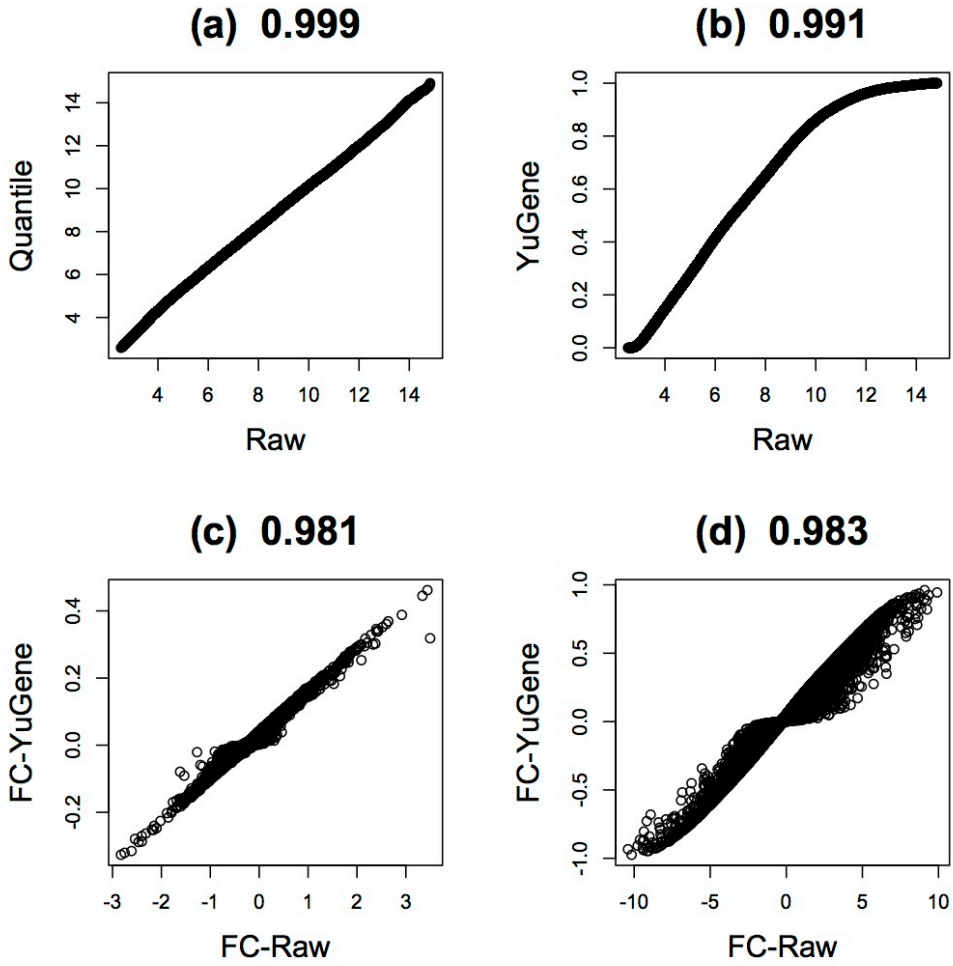
1. Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.M., Goodsaid, F.M., Pusztai, L. *et al.* (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, **28**, 827-838.
2. Mane, S., Evans, C., Cooper, K., Crasta, O., Folkerts, O., Hutchison, S., Harkins, T., Thierry-Mieg, D., Thierry-Mieg, J. and Jensen, R. (2009) Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC genomics*, **10**, 264.
3. Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H. *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J*, **10**, 278-291.
4. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. and Storey, J.D. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*.
5. Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118-127.
6. Bravo, H.C., Pihur, V., McCall, M., Irizarry, R.A. and Leek, J.T. (2012) Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC bioinformatics*, **13**, 272.
7. Piccolo, S.R., Withers, M.R., Francis, O.E., Bild, A.H. and Johnson, W.E. (2013) Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences*, **110**, 17778-17783.
8. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.
9. McCall, M.N., Bolstad, B.M. and Irizarry, R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242-253.
10. McCall, M.N. and Irizarry, R.A. (2011) Thawing Frozen Robust Multi-array Analysis (fRMA). *BMC bioinformatics*, **12**, 369.
11. Turnbull, A., Kitchen, R., Larionov, A., Renshaw, L., Dixon, J. and Sims, A. (2012) Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC medical genomics*, **5**, 35.
12. Cheadle, C., Vawter, M.P., Freed, W.J. and Becker, K.G. (2003) Analysis of microarray data using Z score transformation. *J Mol Diagn*, **5**, 73-81.
13. Reddy, T.B., Riley, R., Wymore, F., Montgomery, P., DeCaprio, D., Engels, R., Gellesch, M., Hubble, J., Jen, D., Jin, H. *et al.* (2009) TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res*, **37**, D499-508.
14. Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, **39**, D1002-D1004.

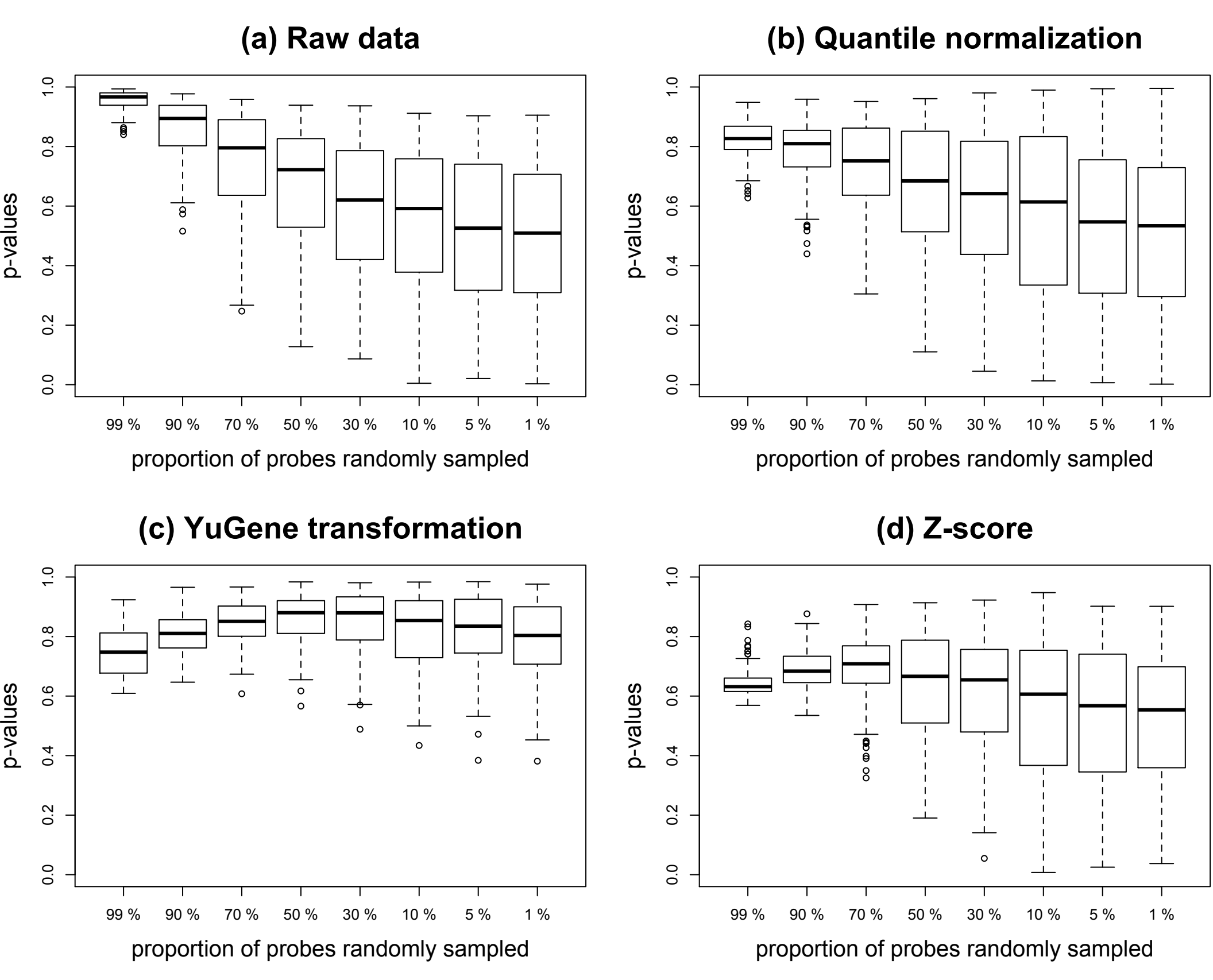
15. Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E. and Brazma, A. (2010) A global map of human gene expression. *Nat Biotech*, **28**, 322-324.
16. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C., Haase, J., Janes, J., Huss, J. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, **10**, R130.
17. Seita, J., Sahoo, D., Rossi, D.J., Bhattacharya, D., Serwold, T., Inlay, M.A., Ehrlich, L.I.R., Fathman, J.W., Dill, D.L. and Weissman, I.L. (2012) Gene Expression Commons: An Open Platform for Absolute Gene Expression Profiling. *PLoS One*, **7**, e40321.
18. Muller, F.J., Schuldt, B.M., Williams, R., Mason, D., Altun, G., Papapetrou, E.P., Danner, S., Goldmann, J.E., Herbst, A., Schmidt, N.O. *et al.* (2011) A bioinformatic assay for pluripotency in human cells. *Nat Methods*, **8**, 315-317.
19. Feingold, E.A., Good, P.J., Guyer, M.S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F.S. and members, E.c. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636-640.
20. Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Carninci, P. and Hayashizaki, Y. (2006) CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res*, **34**, D632-636.
21. Qu, H. and Fang, X. (2013) A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. *Genomics Proteomics Bioinformatics*.
22. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*, **39**, D1005-D1010.
23. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-315.
24. Carvalho, B.S. and Irizarry, R.A. (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, **26**, 2363-2367.
25. Bolstad, B.M. (accessed 2012) preprocessCore: A collection of pre-processing functions. *R packageversion 1.24.0*.
26. Leek, J., Johnson, W., Parker, H., Jaffe, A. and Storey, J. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882 - 883.
27. Team, R.C. (2012), *R Foundation for Statistical Computing*. 2013 ed, Vienna, Austria.
28. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and Core Team, R. (2014) nmle: Linear and nonlinear mixed effects models. http://CRAN.R-project.org/package=nlme., **R package version 3.1-115**.
29. Yoav, B. and Yosef, H. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.
30. Guenther, M.G., Frampton, G.M., Soldner, F., Hockemeyer, D., Mitalipova, M., Jaenisch, R. and Young, R.A. (2010) Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells. *Cell stem cell*, **7**, 249-257.
31. Wells, C., Mosbergen, R., Korn, O., Choi, J., Seidenman, N., Matigian, N.A., Vitale, A.M. and Shepherd, J. (2012) Stemformatics: Visualisation and sharing of stem cell gene expression. *Stem Cell Research*, **10**, 387-395.

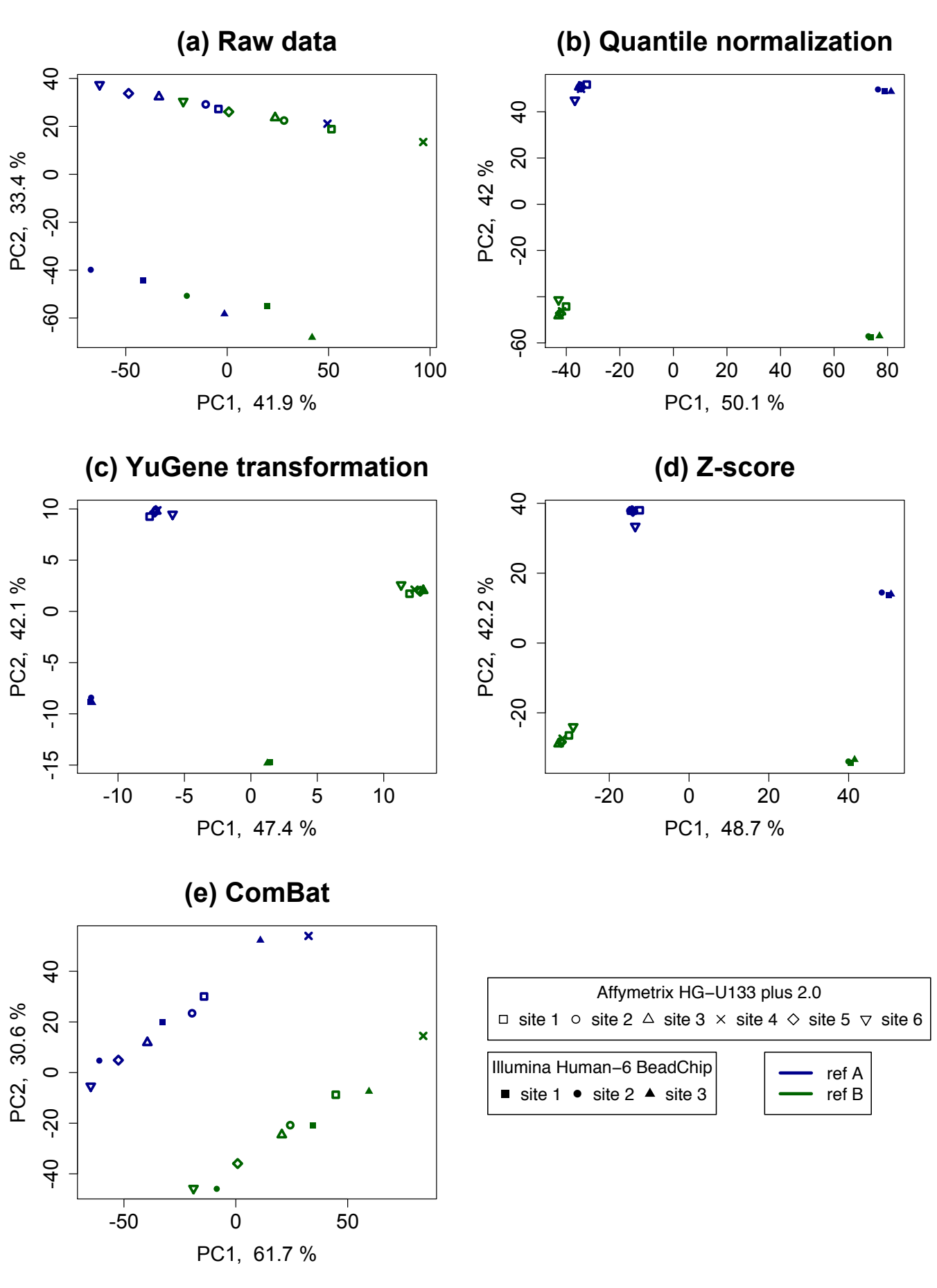
32. McCall, M.N., Uppal, K., Jaffee, H.A., Zilliox, M.J. and Irizarry, R.A. (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res*, **39**, D1011-D1015.
33. Allen, J.D., Wang, S., Chen, M., Girard, L., Minna, J.D., Xie, Y. and Xiao, G. (2011) Probe mapping across multiple microarray platforms. *Briefings in Bioinformatics*.
34. Kuo, W.P., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M., Sarang, J., Whipple, M.E., Maysuria, M., Serikawa, K., Lee, S.Y. *et al.* (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotech*, **24**, 832-840.
35. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.
36. Purdom, E., Simpson, K.M., Robinson, M.D., Conboy, J.G., Lapuk, A.V. and Speed, T.P. (2008) FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, **24**, 1707-1714.
37. Gibson, G. (2010) Hints of hidden heritability in GWAS. *Nature genetics*, **42**, 558-560.
38. Mason, E., Tronc, G., Nones, K., Matigian, N., Kim, J., Aronow, B.J., Wolfinger, R.D., Wells, C. and Gibson, G. (2010) Maternal Influences on the Transmission of Leukocyte Gene Expression Profiles in Population Samples from Brisbane, Australia. *PLoS ONE*, **5**, e14479.
39. Jakubek, Y. and Cutler, D. (2012) A model of binding on DNA microarrays: understanding the combined effect of probe synthesis failure, cross-hybridization, DNA fragmentation and other experimental details of affymetrix arrays. *BMC genomics*, **13**, 737.
40. Skvortsov, D., Abdueva, D., Curtis, C., Schaub, B. and Tavaré, S. (2007) Explaining differences in saturation levels for Affymetrix GeneChip® arrays. *Nucleic Acids Res*, **35**, 4154-4163.
41. Newman, A.M. and Cooper, J.B. (2010) Lab-Specific Gene Expression Signatures in Pluripotent Stem Cells. *Cell stem cell*, **7**, 258-262.
42. Jia, F., Wilson, K.D., Sun, N., Gupta, D.M., Huang, M., Li, Z., Panetta, N.J., Chen, Z.Y., Robbins, R.C., Kay, M.A. *et al.* (2010) A nonviral minicircle vector for deriving human iPS cells. *Nat Meth*, **7**, 197-199.
43. Maherali, N., Ahfeldt, T., Rigamonti, A., Utikal, J., Cowan, C. and Hochedlinger, K. (2008) A High-Efficiency System for the Generation and Study of Human Induced Pluripotent Stem Cells. *Cell stem cell*, **3**, 340-345.
44. Si-Tayeb, K., Noto, F.K., Nagaoka, M., Li, J., Battle, M.A., Duris, C., North, P.E., Dalton, S. and Duncan, S.A. (2010) Highly efficient generation of human hepatocyte–like cells from induced pluripotent stem cells. *Hepatology*, **51**, 297-305.
45. Masaki, H., Ishikawa, T., Takahashi, S., Okumura, M., Sakai, N., Haga, M., Kominami, K., Migita, H., McDonald, F., Shimada, F. *et al.* (2007) Heterogeneity of pluripotent marker gene expression in colonies generated in human iPS cell induction culture. *Stem Cell Research*, **1**, 105-115.

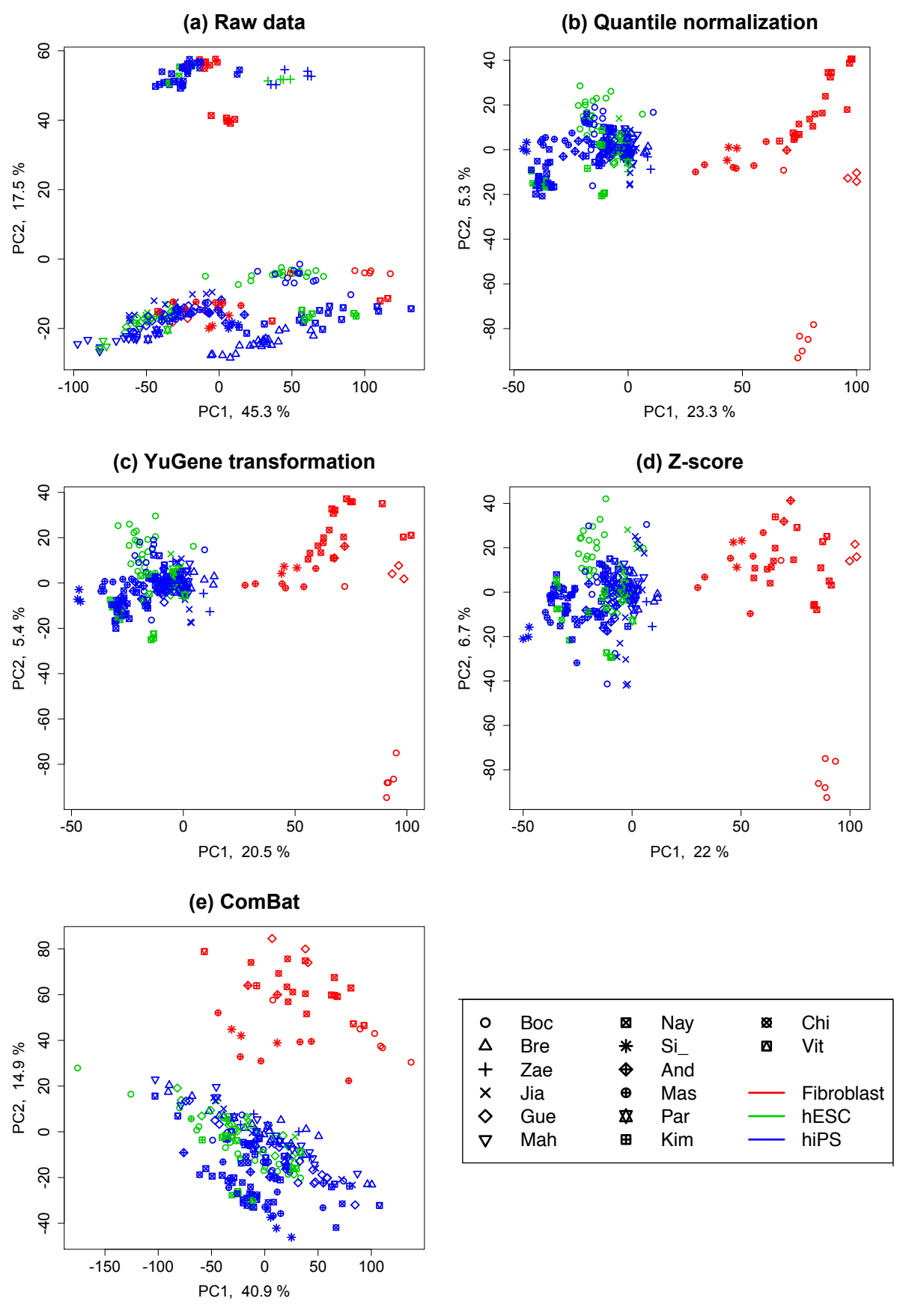
46. Park, I.H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W. and Daley, G.Q. (2008) Reprogramming of human somatic cells to pluripotency with defined factors. *Nature*, **451**, 141-146.
47. Kim, D., Kim, C.-H., Moon, J.-I., Chung, Y.-G., Chang, M.-Y., Han, B.-S., Ko, S., Yang, E., Cha, K.Y., Lanza, R. *et al.* (2009) Generation of Human Induced Pluripotent Stem Cells by Direct Delivery of Reprogramming Proteins. *Cell stem cell*, **4**, 472-476.
48. Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J. *et al.* (2009) Induced Pluripotent Stem Cells and Embryonic Stem Cells Are Distinguished by Gene Expression Signatures. *Cell stem cell*, **5**, 111-123.
49. Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H. *et al.* (2011) Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. *Cell*, **144**, 439-452.
50. Brennand, K.J., Simone, A., Jou, J., Gelboin-Burkhart, C., Tran, N., Sangar, S., Li, Y., Mu, Y., Chen, G., Yu, D. *et al.* (2011) Modelling schizophrenia using human induced pluripotent stem cells. *Nature*, **473**, 221-225.
51. Andrade, L.N.d.S., Nathanson, J.L., Yeo, G.W., Menck, C.F.M. and Muotri, A.R. (2012) Evidence for premature aging due to oxidative stress in iPSCs from Cockayne syndrome. *Human molecular genetics*, **21**, 3825-3834.
52. Zaehres, H., Kögler, G., Arauzo-Bravo, M.J., Bleidissel, M., Santourlidis, S., Weinhold, S., Greber, B., Kim, J.B., Buchheiser, A., Liedtke, S. *et al.* (2010) Induction of pluripotency in human cord blood unrestricted somatic stem cells. *Experimental hematology*, **38**, 809-818.e802.
53. Nayler, S., Gatei, M., Kozlov, S., Gatti, R., Mar, J.C., Wells, C.A., Lavin, M. and Wolvetang, E. (2012) Induced Pluripotent Stem Cells from Ataxia-Telangiectasia Recapitulate the Cellular Phenotype. *Stem Cells Translational Medicine*.
54. Vitale, A.M., Matigian, N.A., Ravishankar, S., Bellette, B., Wood, S.A., Wolvetang, E.J. and Mackay-Sim, A. (2012) Variability in the Generation of Induced Pluripotent Stem Cells: Importance for Disease Modeling. *Stem Cells Translational Medicine*, **1**, 641-650.

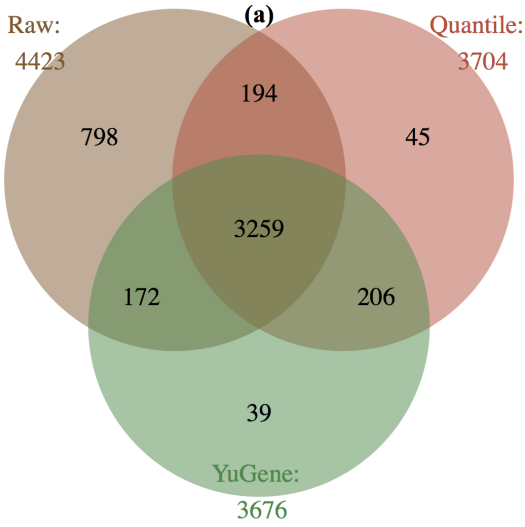

(a) Raw data
density
(b) Quantile normalization
density
(c) YuGene transformation
density
(d) Rank data
density
(e) Z−score
density

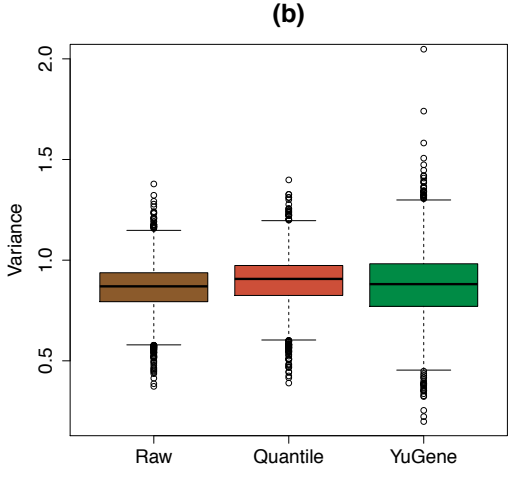

(a) 0.999
Quantile
4
6
8
10
12
14
Raw
(b) 0.991
YuGene
0.0
0.2
0.4
0.6
0.8
1.0
Raw
(c) 0.981
FC-YuGene
-0.2
0.0
0.2
0.4
-3
-2
-1
0
1
2
3
FC-Raw
(d) 0.983
FC-YuGene
-1.0
-0.5
0.0
0.5
1.0
-10
-5
0
5
10
FC-Raw

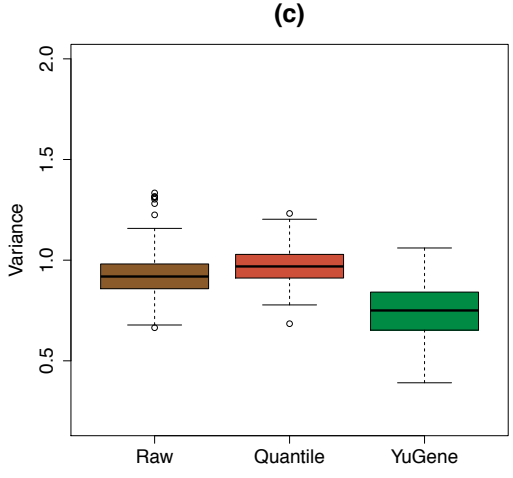

(a) Raw data
(b) Quantile normalization
(c) YuGene transformation
(d) Z-score
p-values
0.0
0.2
0.4
0.6
0.8
1.0
99 %
90 %
70 %
50 %
30 %
10 %
5 %
1 %
proportion of probes randomly sampled


(a) Raw data
PC2, 33.4 %
PC1, 41.9 %
(b) Quantile normalization
PC2, 42 %
PC1, 50.1 %
(c) YuGene transformation
PC2, 42.1 %
PC1, 47.4 %
(d) Z-score
PC2, 42.2 %
PC1, 48.7 %
(e) ComBat
PC2, 30.6 %
PC1, 61.7 %
Affymetrix HG−U133 plus 2.0
site 1 site 2 site 3 site 4 site 5 site 6
Illumina Human−6 BeadChip
site 1 site 2 site 3
ref A
ref B


(a) Raw data
PC2, 17.5 %
PC1, 45.3 %
(b) Quantile normalization
PC2, 5.3 %
PC1, 23.3 %
(c) YuGene transformation
PC2, 5.4 %
PC1, 20.5 %
(d) Z-score
PC2, 6.7 %
PC1, 22 %
(e) ComBat
PC2, 14.9 %
PC1, 40.9 %
Boc
Bre
Zae
Jia
Gue
Mah
Nay
Si_
And
Mas
Par
Kim
Chi
Vit
Fibroblast
hESC
hiPS


(a)
Raw:
4423
Quantile:
3704
YuGene:
3676
798
194
45
3259
172
206
39


(b)
Variance
0.5
1.0
1.5
2.0
Raw
Quantile
YuGene


(c)
Variance
0.5
1.0
1.5
2.0
Raw
Quantile
YuGene